Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 004.9
DOI: https://doi.org/10.24867/13BE31Skiljevic

PREDIKCIJA POZICIJE FUDBALSKOG IGRAČA UPOTREBOM ALGORITAMA MAŠINSKOG UČENJA

PREDICTION OF FOOTBALL PLAYER'S POSITION USING MACHINE LEARNING ALGORITHMS

Dragan Škiljević, *Fakultet tehničkih nauka, Novi Sad*

Oblast – ELEKTROTEHNIKA I RAČUNARSTVO

Kratak sadržaj – *Fudbal je kolektivni sport koji se igra između dvije ekipe, sa po jedanaest igrača. Iako igrači igraju na unaprijed određenoj poziciji, oni mogu lako preći i na neku drugu poziciju. U ovome radu je vršena predikcija najbolje pozicije igrača na osnovu njegovih fizičkih i psihičkih osobina. Osnovni motiv ovoga rada jeste olakšavanje posla fubalskim stručnjacima koji se profesionalno bave svojim poslom. Rješenje ovoga projekta bi u velikoj mjeri olakšalo posao trenerima čiji klubovi se susreću sa mnoštvom povreda, pa je potrebno često vršiti promjenu formacije tima. To bi pomoglo da se u maksimalnoj mjeri iskoristi potencijal svakog igrača. Da bi se što lakše odredila pozicija na kojoj će određeni igrač igrati, u ovom radu, koristićemo skup podataka sa 65 atributa za svakog igrača, na osnovu kojih će se određivati pozicija uz pomoć obučavanja sledećih modela: Multiomial Logistic Regression, K-Nearest Neighbors, Random Forest, Gaussian Naive Bayes, Suport Vector Machine.*

Ključne reči: *fudbal, pozicija, mašinsko učenje, klasifikacija, multilabel*

Abstract – *Football is a collective sport that is played between two teams, composed of eleven players each. Although players play in a predetermined position, they can quickly move to another position. In this paper, the prediction of the best position for a football player was made based on his physical and mental characteristics. This approach's primary motivation is to provide additional information to coaches whose clubs are faced with many injured players. Each player was represented with a vector of 65 attributes, and the following models: Multinomial Logistic Regression, K-Nearest Neighbors, Random Forest, Gaussian Naive Bayes, Support Vector Machine were experimented with.*

Keywords: *football, position, machine learning, classification, multilabel*

1. UVOD

S obzirom da je fudbal najvažnija sporedna stvar na svijetu, on je široko prisutan u svim narodnim masama. Sa porastom popularnosti rastu i cijene igrača koje u današnje vrijeme dostižu milionske iznose.

Cijene igrača su takođe tijesno povezane sa pozicijom na kojoj igrač igra. Obično najveće cijene imaju fudbaleri koji igraju u napadu. Kako se sve veći novac ulaže u kupovinu igrača, veoma je bitno da se uloženi novac isplati. Pozicija na kojoj će igrač igrati je jedan od ključnih faktora da bi se iskoristio sav njegov potencijal.

Poznato je da ishod fudbalske utakmice često zavisi od izbora pravilne formacije. Kako je izbor formcije lična odluka trenera, tako je on suočen sa problemom izbora najboljih igrača za svaku poziciju. Cilj ovog rada je izbor najbolje pozicije za fudbalskog igrača na osnovu njegovi fizičkih i psihičkih osobina. To bi u velikoj mjeri pomoglo trenerima prilikom izbora igrača za određenu poziciju. Sistem predstavljen u ovom radu bi u velikoj mjeri eliminisao subjektivnu pristrasnost prilikom izbora igrača koji će nastupiti. To bi kasnije dovelo do boljih rezultata i smanjenja nepotizma u timu. Fudbalski skauti su često izloženi problemu izbora igrača koji bi se najbolje uklopilo u njihov tim. Pošto su cijene igrača enormno porasle pravljenje grešaka mora biti svedeno na minimum, jer svaka greška može da dovede do finansijske krize kluba. Ovo rješenje bi takođe olakšalo rješavanje ovog problema.

U ovom radu će biti predstavljeno rješenje koje će vršiti izbor pozicije na osnovu sledećih parametara: brzina, kontrola lopte, dribling, skok, pregled igre, agresivnost, smirenost itd.

Preostali dio rada ogranizovan je na sledeći način. U drugom poglavlju će biti prezentovani radovi koji se bave sličnom problematikom. U trećem poglavlju će biti opisan skup podataka i način pripreme podataka za obučavanje i validaciju modela. U četrvtom poglavlju biće predstavljena metodologija za klasifikaciju koja je korišćena za rešavanje problema određivanja najbolje pozicije igrača. U petom poglavlju su prikazani i diskutovani rezultati koji su postigli modeli. Na kraju će biti izveden zaključak ovog rada i navedeni pravci u kojima bi se dalje mogla razvijati i unapređivati ova problematika.

2. PRETHODNA RJEŠENJA

U radu [1], koji se bavi predviđanjem i prepoznavanjem talenta u fudbalu na osnovu individualnih kvaliteta igrača. Kvaliteti su klasifikovani u 3 grupe: fizičke, mentalne i tehničke. Tu spadaju: brzina, agilnost, skok, visina, snaga, sposobnost čitanja igre, smirenost, kreativnost, samouvjerenost, sposobnost u predvođenju tima, pas igra, šut, igra glavom, završnica, uklizavanje, ubačaj, odbrana.

NAPOMENA:
Ovaj rad proistekao je iz master rada čiji mentor je bio dr Aleksandar Kovačević, vanr. prof.

Algoritmi klasifikacije koji su korišteni u ovom radu su: *Bayesian Networks*, *Decision Trees*, i *K-Nearest Neighbor*. Najveća tačnost klasifikacije postignuta tokom eksperimenata iznosila je 99% za *Bayesian Networks*, *Decision Tree* 98%, a *Nearest Neighbor* 97%. Dodatna evaluacija sistema je vršena uz pomoć 20 fudbalskih strucnjaka koji su bili treneri i/ili fudbalski menadžeri. Od njih je zatrazeno da prodju kroz funkcionalnosti koje sistem nudi i ostave povratnu informaciju o tome kako su zadovoljni sistemom.Analizom je utvrđeno da se 70% korisnika u potpunosti slaže i misli da bi sistem bio od pomoći u realnom svijetu, 15% se prilicno slaže sa sistemom, a 15% korisnika nisu sigurni da bi sistem bio od koristi i da bi se mogao implementirati u fudbalskim organizacijama.

Cilj projekta [2] jeste implementacija modela za preporučivanje pozicije igrača na osnovu fizičkih atributa. Rješenje je predstavljeno uz pomoć sledećih linearnih tehnika: *Linear regression*, *Linear discriminant analysis*, *Quadratic discriminant analysis* i *Multinomial logistic regression*. Ulazne promjenljive su snaga i izdržljivost i imaju vrijednosti između 0 i 100. Izlazna promjenljiva je pozicija i uzima vrijednosti {"CB": srednji bek, "CM": srednji vezni, "ST": napadač} koji predstavljaju 3 ključne pozicije. Odbrambeni igrači su klasifikovani tako da su u prosjeku jači i imaju manje izdržljivosti. Igrači centralnog veznog reda su u prosjeku prepoznati kao slabiji. Karakteristike napadača su dobra kontrola lopte. Prema podacima, prosječna snaga i izdržljivost napadača su ustvari između prosjeka za odbrambene i centralne vezne igrače. Rezultati su dobijeni uz pomoć *precision* metrike.

LR = 50.5% (CB: 52.9% CM: 56.6%, ST: 41.9%)

LDA = 51.0% (CB: 54.5%, CM: 57.0%, ST: 41.5%)

QDA = 51.3% (CB: 53.1%, CM: 57.4%, ST: 43.4%)

MLR = 51.1% (CB: 54.2%, CM: 57.4%, ST: 41.7%)

Za razliku od navedenog rada, dodate su nove pozicije: golman, bek i krilo.

Rad [3] se bavi istraživačkom analizom uz pomoć koje se predviđa položaj odnosno pozicija igrača koristeći različite algoritme mašinskog učenja. Podaci koju su korišćeni sadrže oko 18.000 igrača sa 75 karakteristika po igraču. Prilikom izrade modela korišteni su sledeći algoritmi: *K-Nearest Neighbors*, *Random Forest*, *Support Vector Machine*, *Neural Network* (NNET). Tačnost pomenutih algoritama iznosila je: KNN-81.97%; RF-83.06%; SVM-82.89%; NNET 0. 00%. Ovo nije iznenađenje jer se njihove tehničke osobine znatno razlikuju. Bitno je napomenuti da je tačnost prilikom izbora golmana iznosila 100%. Za razliku od pomenutog rada, sistem predstavljen u ovom radu omogućava da igrač može igrati na više od jedne pozicije, što dovodi do *multilabel* problema koji ćemo riješiti na nekolika načina. Prilikom izrade modela iskoristićemo gore pomenute algoritme Random Forest i SVM.

Rad [4] je vršio je predikciju pozicije igrača iz popularne igre FIFA 19. Skup podataka je preuzet iz FIFA 19 baze podataka koja sadrži 18207 igrača od kojih svaki sadrži preko 80 različitih atributa. U inicijalnom skupu podataka igračima je bila dodjeljena 1 od 27 pozicija. Da bi se uprostila struktura formirane su ukupno 4 pozicije koje mogu biti dodjeljene igračima i to su: golman, odbrambeni, vezni, i napadač.Prilikom izrade modela za predikciju pozicije igrača su korišteni *Decision Tree* i *Random Forest* algoritmi. Koristeći *Decision Tree* algoritam postignuta je tačnost od 85% uz pomoć *accuracy* metrike koja predstavlja odnos tačno klasifikovanih igrača i ukupnog broja igrača. Tačnost po klasama je bila najveća za golmana i iznosila je 100%. Odbrambeni igrači su tačno klasifikovani u 79% slučajeva, vezni igrači u 84% slučajeva, a napadači su ostvarili tačnost od 91%. Koristeći *Random Forest* algoritmam ostvarena je tačnost od 93%. Analizom atributa izabran je najuticajniji atribut koji je u ovom slučaju bio *sliding tackle* (uklizavanje). Samo na osnovu ovoga parametra moguće je odrediti poziciju sa tačnošću od 72%.

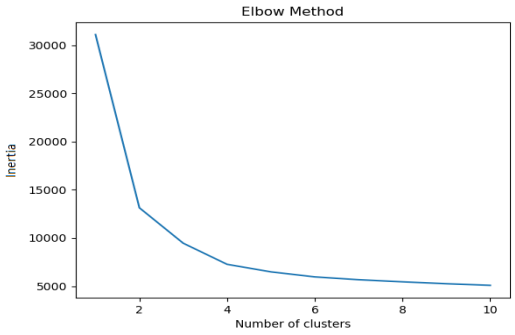
3. SKUP PODATAKA

U ovom poglavlju se opisuje postupak koji dovodi do formiranja setova podataka, koji se dalje koriste za formiranje modela sistema. Inicijalni set podataka je preuzet sa stranice (https://public.tableau.com/s/sites/default/ files/media/fifa18_cl ean.csv), a evaluacija parametara je vršena na osnovu svjetskog prvenstva 2018. u Rusiji. Skup podataka se sastoji od 17981 igrača od kojih svaki posjeduje 65 parametara. Neki od glavnih parametara su: godine igrača, brzina, agresivnost, agilnost, kontrola lopte, smirenost, ubačaj, driblig, volej, uklizavanje, igra glavom, skok, sprint, snaga, izvođenje penala, pregled igre, itd.

Ovaj skup podataka je podijeljen u omjeru 80 : 10 : 10, što znači da se 80% podataka koristi kao obučavajući skup, 10% kao validacioni skup, a preostalih 10% kao testni skup.

3.1 Analiza podataka

Inicijalni data set posjeduje 12 pozicija na kojima igrači mogu biti raspoređeni. Analizom klastera uz pomoć *Elbow* metode dobijeni su rezultati prikazani na gafiku 3.2.1.

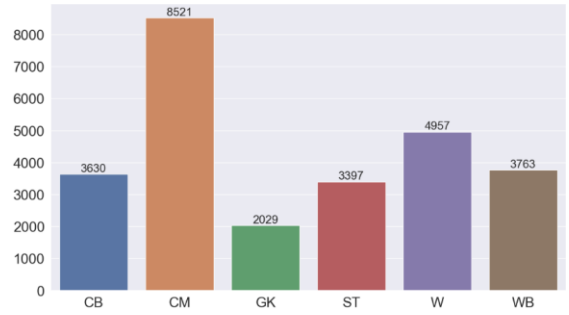


Grafik 3.2.1 *Rezultat Elbow metode*

Sa slike se vidi da 12 ne predstavlja optimalan broj klasa, već je optimum 4, ali s obzirom da današnji fudbal zahtjeva više od 4 pozicije, kao kompromis sistem će koristiti 6 pozicija nad kojima će se izvršavati algoritmi, a to su: Golman (GK), Štoper (CB), Bek (WB), Vezni (CM), Krilo (W), Špic (ST).

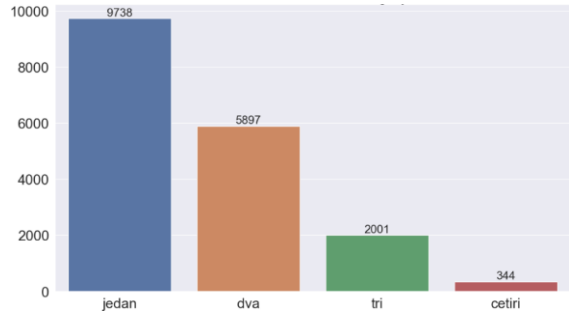
Kako inicijalni skup podataka sadrži neke parametre koji nisu potredni za analizu, kao sto su ime, klub, država, mjesto rođenja itd. oni su izbačeni iz skupa atributa. Golman posjeduje neke atribute koji nemaju vrijednosti, pa su u cilju smanjenja grešaka ove vrijednosti popunjene nulama.

Daljom analizom inicijalnog skupa podataka je utvrđen broj igrača koji igraju na određenim pozicijama. Rezultat analize je prikazan na grafiku 3.2.1



Grafik 3.2.1 *Raspodjela podataka po pozicijama*

Sa slike se može jasno vidjeti da najveći dio igrača iz skupa podataka pripada kategoriji veznih igrača, jer se oni ne previše ne ističu ni po kojim osobinama i predstavljaju najveći problem prilikom klasifikacije. Nasuprot tome igrači na poziciji golmana, koji imaju istaknute odbrambene osobine, predstavljaju najmanji dio data seta, ali se jasno razlikuju po karakteristikama. S obzirom da se vidi sa slike da je zbir svih igrača po pozicijama veći od ukupnog broja igrača iz data seta, dolazimo do zaključka da neki igrači mogu igrati na više pozicija. Daljom analizom je utvrđena podjela igrača po broju pozicija, a to je prikazano na grafiku 3.2.3.



Grafik 3.2.3 *Raspodjela podataka u odnosu na broj pozicija*

Grafik 3.2.3 nam pokazuje jasno da najveći broj igrača igra na samo jednoj poziciji odnosno 9738 igrača što čini 54.15% od ukupnog broja.

U inicijalnom setu postoje i neka obilježja čije vrijednosti nedostaju pa su, zbog nemogućnosti pojedinih algoritama da formiraju model nad nepotpunim skupom podataka, te vrijednosti zamijenjene nulom.

Normalizacija podataka jedan je od pristupa pre obrade gdje se podaci skaliraju ili transformiraju kako bi dali jednak doprinos svakog obilježja. Kako u ovome slučaju postoje obilježja čiji se domeni u velikoj mjeri razlikuju, prije pravljenja modela za predikciju neophodno je obilježja normalizovati. U ovome slučaju normalizacija je vršena po sledecoj formuli:

$$X' = \frac{X}{X_{max}}$$

S obzirom da naša glavna labela Pozicija predstavlja kategoričko obilježje tj. sadrži konačan skup diskretnih vrijednosti bez međuzavisnosti između njih, ona se kao takva mora transformisati u drugi oblik da bi se prilagodila većini algoritama mašinskog učenja. Biće korištena *One Hot Encoding* transformacija koja od obilježja formira matricu nula i stavlja jedinice na mjesto za onu kolonu, čiju vrijednost posjeduje.

Daljom analizom data seta ispitane su zavisnosti među atributima. To je postignuto uz pomoć matrice korelacija na osnovu koje je utvrđeno da postoje atributi koji su u jakoj korelacionoj vezi. To predstavlja jasan razlog za izbacivanje jednog od koreliranih atributa kako bi se smanjila dimenzionalnost i složenost modela, a samim tim i smanjilo vrijeme potrebno za predikciju. Matricom korelacija je utvrđeno da su atributi GK_reflexes i GK_positioning nalaze u korelaciji, te je prvi atribut izbačen iz skupa podataka.

4. METODOLOGIJA

Algoritmi koji su korišćeni i analizirani u ovom radu su: *Naive Bayes, Support Vector Machines (SVM), Random Forest (RF), K Nearest Neighbors (KNN),Multinomial Logistic Regresion (MLR).*

Svaki klasifikator je implementiran na određeni način i nakon toga je vršena evaluacija dobijenih rezultata. Metrike evaluacije performansi klasifikatora koje su korištene u ovom radu su: *Hamming loss, Jaccard score, F1 score.*

Kao optimalni kernel za SVM algoritam je izabran *rbf*, a paramer regularizacije C je određen empirijskim putem i iznosi 1. Za RF algoritam optimalna vrijednost parametra n_estimators, koji predstavlja broj stabala koja će se formirati unutar algoritma, iznosi 250. Za KNN algoritam optimalna vrijednost parametra n_neighbors iznosi 11.

Multilabel problem će biti rešavan na 2 načina, pomoću *Classifier Chains (CC)* algoritma i *Binary Relevance (BR)* algoritma i biće prikazani rezultati u kombinaciji sa svim klasifikacionim algoritmima.

5. ANALIZA REZULTATA I DISKUSIJA

U ovom poglavlju će biti predstavljeni rezultati dobijeni primjenom svakog od ponuđenih algoritama i diskusija njihovih rezultata.

Tabela 1. *Rezultati SVM algoritma*

SVM	BR	CC
F1 score	85.18%	84.38%
Hamming loss	0.05	0.06
Jaccard index	83.56%	82.94%

SVM algoritam je opravdao svoju popularnost i pokazao se kao klasifikator koji je donio najbolje rezultate. Rešenje uz pomoć BR metode je imalo tačnost od 85.18%, dok je rešenje uz pomoć CC metode imalo tačnost od 84.38%. Ostvareni rezultat je gotovo jednak rezultatu u radu [3] koji iznosi 82.89%, tačnije razlika je oko 2% što je u našem slučaju zanemarivo.

Tabela 2. *Rezultati Naive Bayes algoritma*

Naive Bayes	BR	CC
F1 score	63.85%	62.93%
Hamming Loss	0.21	0.22
Jaccard index	52.38%	52.65%

Naive Bayes algoritam je svoju najbolju tačnost pokazao zajedno sa BR algoritmom i ona iznosi 63.85%. Nešto manju tačnost za oko 1% ima CC algoritam i ona iznosi 62.93%. Ipak ovaj algoritam nije pokazao izrazito dobre rezultate u odnosu na ostale algoritme, ali ono u čemu se on izdvaja jeste vrijeme njegovog izvršavanja koje iznosi oko 0.5 sekundi, što ga stavlja na prvu poziciju po brzini izvršavanja.

Tabela 3. *Rezultati Random Forest algoritma*

RF	BR	CC
F1 score	84.75%	85.06%
Hamming Loss	0.06	0.05
Jaccard index	83.17%	83.53%

Random Forest algoritam je pokazao veoma dobru tačnost prilikom predviđanja i ona iznosi 85.06% za standardne parametre u kombinaciji sa CC algoritmom. Za oko 0.30% manje rezultate nudi BR metoda, a njena tačnost jeste 84.75%. U radu [4] se ovaj algoritam pokazao kao najbolje rešenje sa tačnošću od 93%, međutim u ovom radu je broj pozicija veći od 4, što je dovelo do smanjenja tačnosti rezultata. RF algoritam je dao rezultate slične SVM algoritmu kao i u radu [3]. Ono što je bitno napomenuti jeste da ovaj algoritam ima najduže vrijeme izvršavanja i da se ono proporcionalno povećava u odnosu na broj stabala formiranih u okviru RF algoritma.

Tabela 4. *Rezultati KNN algoritma*

Metrike	BR	CC
F1 score	82.50%	82.51%
Hamming Loss	0.07	0.06
Jaccard index	80.67%	80.97%

KNN algoritam je dao veoma dobre rezultate u odnosu na svoju jednostavnost. Rezultati su bili veoma slični i za CC i za BR algoritam i iznosili su respektivno 82.51% i 82.50%. Neznatno malu prednost ima CC algoritam i to za 0.01%. Slično kao u radovima [1] i [3] ovaj algoritam se nije pokazao kao najbolje rešenje, ali veoma malo zaostaje u odnostu na optimalno rešenje. Ono što je interesantno za ovaj algoritam jeste relativno malo vrijeme obučavanja modela ali veliko vrijeme predviđanja koje je ponekad veće od vremena obučavanja. Analizom atributa izabrano je 50% najuticajnih atributa koji su ostvarili veću tačnost za 0.5%.

Tabela 5. *Rezultati MLR algoritma*

Metrike	BR	CC
F1 score	83.16%	82.96%
Hamming Loss	0.07	0.07
Jaccard index	80.51%	80.56%

MLR klasifikacioni algoritam ponudio je veoma dobre rezultate uzimajući u obzir njegovo malo vrijeme izvršavanja, a po tom parametru se nalazi odmah iza *Naive Bayes* algoritma. Najveću tačnost je dao zajedno sa BR algoritmom i ona iznosi 83.16%. Nešto manju tačnost, za oko 0.20% dao je CC algoritam za rešavanje multilabel problema, a ona iznosi 82.96%. U radu [3] ostvaren je dosta slabiji rezultat od 51.1%, ali je za evaluaciju rezultata korištena *precision* metrika, što predstavlja veoma bitnu razliku. Analizom izbora atributa ostvareni su rezultati koji nisu imali bolju tačnost.

6. ZAKLJUČAK

U ovom radu istražena je primjena algoritama mašinskog učenja u procesu klasifikacije podataka koji su prikupljeni sa interneta. Motiv za obradu ove teme prije svega nalazi se u potrebi za određivanjem za koju poziciju fudbaler ima najbolje predispozicije, što bi napravilo ogromnu olakšicu svim trenerima. Obučavajući skup podataka se sastojao iz 17981 igrača sa svojim psihičkim i fizičkim osobinama koje su precjenjene na osnovi stručnjaka iz te oblasti. Procjene su vršene na osnovu svjetskog prvenstva u Rusiji 2018 godine. U ovom slučaju za klasifikaciju iskorišteni su *K-Nearest Neighbours*, *Random Forest, Suport Vector Machine*, *Gaussian Naive Bayes* i *Multinominal Logistic Regression.* Kao tehnike za rešavanje multilabel problema korišteni su *Binary Relevace* i *Classifier Chains* algoritmi. *Classifier chains* algoritam je pokazao nešto bolju tačnost u odnosu na *Binary Relevance* algoritam.

Najbolju tačnost je pokazao SVM klasifikator u kombinaciji sa *Classifier Chains* algoritmom i ona iznosi 85.18%. Osnovna prednost ovoga rada u odnosu na druge radove jeste proširenje broja pozicija uz minimalno gubljenje na tačnosti. Mana ovoga rada jeste to što su svi parametri za igrače preuzeti sa svjetskog prvenstva u Rusiji, što može dovesti do potencijalno pogrešnih parametara jer su oni preuzeti na osnovu trenutnog stanja igrača a ne višegodišnje statistike. Ovaj rad bi mogao da se unaprijedi u više pravaca, recimo da se proširi broj pozicija na osam ili devet koje su zastupljene u današnjem fudbalskom svijetu. S obzirom da je mašinsko učenje dosta popularno u današnje vrijeme smatra se da je njegovo povezivanje sa najpopularnijm sportom ostvarilo dobar rezultat.

7. LITERATURA

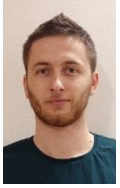
[1] N. Razali, A. Mustapha, F. A. Yatim, „Predicting Player Position for Talent Identification in Association Football", International Research and Innovation Summit (IRIS2017) 6–7 May 2017, Melaka, Malaysia

[2] N. Chandarana,"Football Player Position Prediction", Towards data science, May 2019, Dostupno: https://towardsdatascience.com [Pristupljeno: februar 2021]

[3] D. Schoch, "Predicting player positions", Schocastics, Nov 2017, Dostupno: http://blog.schochastics.net [Pristupljeno: februar 2021]

[4] M. Wiseman "Machine Learning using FIFA 2019", Linkedin, Feb 2019, Dostupno: https://www.linkedin.com [Pristupljeno: februar 2021]

Kratka biografija:



Dragan Škiljević rođen je u Gacku 1996. godine. Master rad na Fakultetu tehničkih nauka iz oblasti Elektrotehnike i računarstva – Računarstvo i automatika odbranio je 2021. godine.

Kontakt: jasamdragan@gmail.com